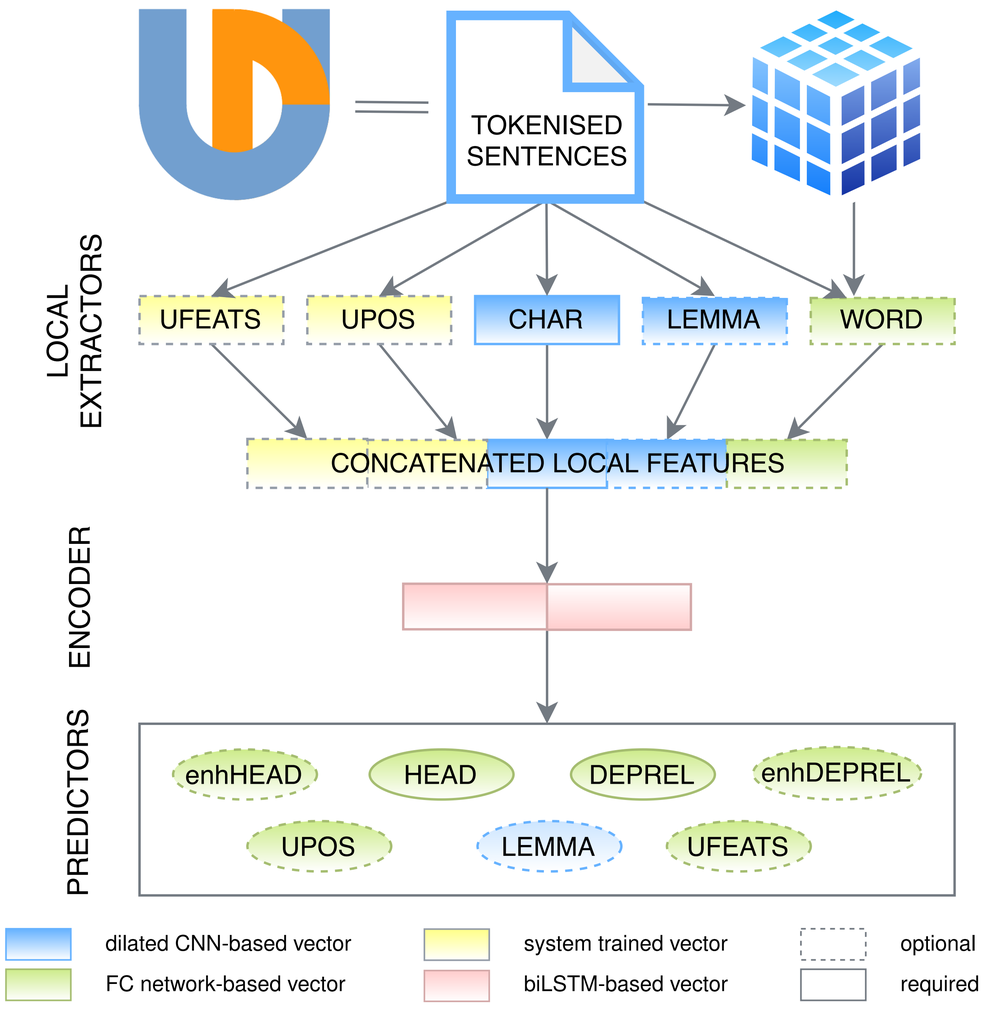
COMBO: State-of-the-Art Morphosyntactic Analysis

Mateusz Klimaszewski[1,2] Alina Wróblewska[2]

m.klimaszewski@ii.pw.edu.pl alina@ipipan.waw.pl

[1]Warsaw University of Technology [2]Institute of Computer Science, Polish Academy of Sciences

COMBO architecture



Qualitative evaluation (F_1 scores)

System	UPOS	XPOS	UFeat	Lemma	UAS	LAS
UD **English** EWT (isolating)						
spaCy	93.79	93.10	94.89	NA	83.38	79.76
Stanza	96.36	96.15	97.01	**98.18**	89.64	86.89
COMBO	**96.57**	**96.44**	**97.24**	97.86	**91.76**	**89.28**
UD **Arabic** PADT (fusional)						
spaCy	90.27	82.15	82.70	NA	74.24	67.28
Stanza	96.98	93.97	94.08	**95.26**	87.96	83.74
COMBO	**97.04**	**94.83**	**95.05**	93.95	**89.21**	**85.09**
UD **Polish** PDB (fusional)						
spaCy	96.14	86.94	87.41	NA	86.73	82.06
Stanza	98.47	94.20	94.42	97.43	93.15	90.84
COMBO	**98.97**	**96.54**	**96.80**	**98.06**	**95.60**	**93.93**
UD **Finnish** TDT (agglutinative)						
spaCy	92.15	93.34	87.89	NA	80.06	74.75
Stanza	97.24	97.96	95.58	**95.24**	89.57	87.14
COMBO	**98.29**	**99.00**	**97.30**	89.48	**94.11**	**92.52**
UD **Korean** Kaist (agglutinative)						
spaCy	85.21	72.33	NA	NA	76.15	68.13
Stanza	95.45	**86.31**	NA	**93.02**	88.42	86.39
COMBO	**95.89**	85.16	NA	89.95	**89.77**	**87.83**

COMBO system

- functional:
  - fully neural
  - predicts categorical morphosyntactic features and exposes their vectors
  - easy to install Python package that uses the PyTorch and AllenNLP libraries
- multipurpose:
  - part-of-speech tagging
  - morphological analysis
  - lemmatisation
  - (enhanced) dependency parsing
- accessible:
  - open-source code (GNU GPL v3.0)
  - pre-trained models for over 40 languages
- efficient:
  - fast model training
  - state-of-the-art prediction quality.

Downstream evaluation (accuracy)

- Task: textual entailment in English and Polish
- Entailment classifier: neural network with two FC layers
- Tested sentence representations:
  - max/mean-pooled BERT embeddings (baseline)
  - morphosyntactically informed BERT embeddings transformed by a network with two transformer layers

Language	max-pooled	mean-pooled	COMBO-based
English	58.1	73.4	78.8
Polish	74.4	83.9	91.6

Efficiency evaluation – training time

Treebank	spaCy	Stanza				COMBO
		Tagger	Lemmatiser	Parser	Total	
English EWT	00:22:34	02:08:51	02:12:17	02:29:13	06:50:21	1:54:11
Polish PDB	01:07:55	04:36:51	03:19:04	05:08:41	13:04:36	3:31:41

Getting started with COMBO

Python

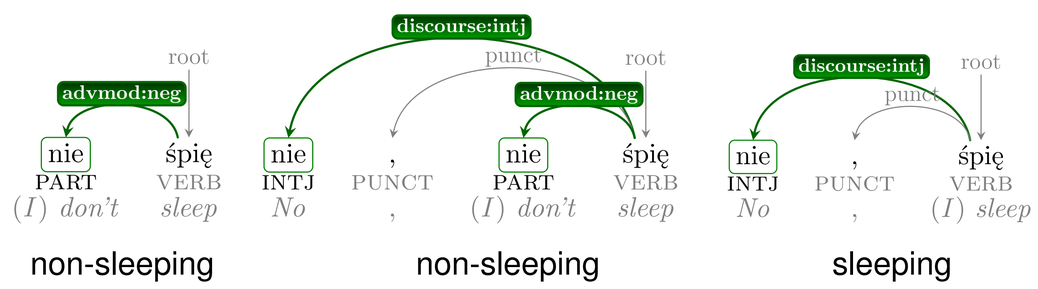
```
from combo.predict import COMBO

nlp = COMBO.from_pretrained("polish")
sentence = nlp("Ala ma kota.")
print(sentence.tokens)
print(sentence.tokens[1].embeddings['upostag'])
```

CLI

```
combo --mode predict \
      --model_path model.tar.gz \
      --input_file input.conllu \
      --output_file output.conllu
```

Why do we need morphosyntactic features?



The authors would like to thank Piotr Rybak for his design and explanations of the architecture of COMBO's forerunner. The research presented in this paper was founded by SONATA 8 grant no 2014/15/D/HS2/03486 from the National Science Centre Poland and the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

Code available at: